

CONDITIONAL QUANTILE ESTIMATION THROUGH OPTIMAL QUANTIZATION

ISABELLE CHARLIER^{(1,2,3)*} DAVY PAINDAVEINE^{(1,2)†} JÉRÔME SARACCO⁽³⁾

May 13, 2014

⁽¹⁾ *Université Libre de Bruxelles, Département de Mathématique, Boulevard du Triomphe, Campus Plaine, CP210, B-1050, Bruxelles, Belgique.*

`ischarli@ulb.ac.be, dpaindav@ulb.ac.be`

⁽²⁾ *ECARES, 50 Avenue F.D. Roosevelt, CP114/04, B-1050, Bruxelles, Belgique.*

⁽³⁾ *Université de Bordeaux, Institut de Mathématiques de Bordeaux, UMR CNRS 5251 et INRIA Bordeaux Sud-Ouest, équipe CQFD, 351 Cours de la Libération, 33405 Talence.*

`Jerome.Saracco@math.u-bordeaux1.fr`

Abstract

In this paper, we use quantization to construct a nonparametric estimator of conditional quantiles of a scalar response Y given a d -dimensional vector of covariates X . First we focus on the population level and show how optimal quantization of X , which consists in discretizing X by projecting it on an appropriate grid of N points, allows to approximate conditional quantiles of Y given X . We show that this approximation is arbitrarily good as N goes to infinity and provide a rate of convergence for the approximation error. Then we turn to the sample case and define an estimator of conditional quantiles based on quantization ideas. We prove that this estimator is consistent for its fixed- N population counterpart. The results are illustrated on a numerical example. Dominance of our estimators over local constant/linear ones and nearest neighbor ones is demonstrated through extensive simulations in the companion paper [Charlier et al. \(2014\)](#).

*Research is supported by a Bourse F.R.I.A. of the Fonds National de la Recherche Scientifique, Communauté française de Belgique.

†Research is supported by an A.R.C. contract from the Communauté Française de Belgique and by the IAP research network grant P7/06 of the Belgian government (Belgian Science Policy).

1 Introduction

In numerous applications, one considers regression modelling to assess the impact of a d -dimensional vector of covariates X on a scalar response variable Y . It is then classical to consider the conditional mean and variance functions

$$x \mapsto \mathbb{E}[Y|X = x] \quad \text{and} \quad x \mapsto \text{Var}[Y|X = x], \quad (1.1)$$

respectively. A much more thorough picture, however, is obtained by considering, for various $\alpha \in (0, 1)$, the conditional quantile functions

$$x \mapsto q_\alpha(x) = \inf \{y \in \mathbb{R} : F(y|x) \geq \alpha\}, \quad (1.2)$$

where $F(\cdot|x)$ denotes the conditional distribution of Y given $X = x$. These conditional quantile functions completely characterize the conditional distribution of Y given X , whereas (1.1), in contrast, only measures the impact of X on Y 's location and scale, hence may completely miss to capture a possible impact of X on the shape of Y 's distribution, for instance.

An important application of conditional quantiles is that they provide reference curves or surfaces (the graphs of $x \mapsto q_\alpha(x)$ for various α) and conditional prediction intervals (intervals of the form $I_\alpha(x) = [q_\alpha(x), q_{1-\alpha}(x)]$, for fixed x) that are widely used in many different areas. In medicine, reference growth curves for children's height and weight as a function of age are considered. Reference curves are also of high interest in economics (e.g., to study discrimination effects and trends in income inequality), in ecology (to observe how some covariates can affect limiting sustainable population size), and in lifetime analysis (to assess influence of risk factors on survival curves), among many others.

Quantile regression, that concerns the estimation of conditional quantile curves, was introduced in the seminal paper [Koenker and Bassett \(1978\)](#), where the focus was on linear regression. Since then, there has been much research on quantile regression, in particular in the nonparametric regression framework. Kernel and nearest-neighbor estimators of conditional quantiles were investigated in [Bhattacharya and Gangopadhyay \(1990\)](#), while [Yu and Jones \(1998\)](#) focused on local linear quantile regression and double-kernel approaches. Many other estimators were also considered; see, among others, [Fan et al. \(1994\)](#), [Gannoun et al. \(2002\)](#), [Heagerty and Pepe \(1999\)](#), or [Yu et al. \(2003\)](#). In this work, we introduce a new nonparametric regression quantile method, based on *optimal quantization*.

Optimal quantization is a tool that was first used by engineers in signal and information theory, where "quantization" refers to the discretization of a continuous signal using a finite number of points, called *quantizers*. The aim being to achieve an efficient, parsimonious, transmission of the signal, the number and location of the quantizers have to be optimized. Quantization was

later used in cluster analysis, pattern and speech recognition. More recently, it was considered in probability theory; see, e.g., [Zador \(1964\)](#) or [Pagès \(1998\)](#). In this context, the problem of optimal quantization consists in finding the best approximation of a continuous d -dimensional probability distribution P by a discrete probability distribution charging a fixed number N of points. In other words, the d -dimensional random vector X needs to be approximated by a random vector \tilde{X}^N that may assume at most N values. Quantization was extensively investigated in numerical probability, finance, stochastic processes, and numerical integration ; see, e.g., [Pagès et al. \(2004a\)](#), [Pagès et al. \(2004b\)](#), or [Bally et al. \(2005\)](#).

Quantization, however, was seldom used in statistics. To the best of our knowledge, its applications in statistics are restricted to Sliced Inverse Regression ([Azaïs et al., 2012](#)) and clustering ([Fischer, 2010](#) and [Fischer, 2013](#)). As announced above, we use in this paper quantization in a nonparametric quantile regression framework. In this context, indeed, quantization naturally takes care of the localization-in- x required in any nonparametric regression method. The resulting quantization-based estimators inherently are based on adaptive bandwidths, hence may dominate the local constant and local linear estimators [Yu and Jones \(1998\)](#) that typically involve a unique global bandwidth. Quantization-based estimators also provide a refinement over nearest-neighbor estimators (such as those from [Bhattacharya and Gangopadhyay \(1990\)](#)) since, unlike the latter, the number of “neighbors” the former consider depends on the point x at which $q_\alpha(x)$ is to be estimated. This dominance over these two standard competitors, in terms of MSEs, is demonstrated through extensive simulations in the companion paper [Charlier et al. \(2014\)](#).

The outline of the paper, that mostly focuses on theoretical aspects, is as follows. Section 2 discusses quantization and provides some results on quantization, both of a theoretical and algorithmic nature. Section 3 describes how to approximate conditional quantiles through optimal quantization, which is achieved by replacing X in the definition of conditional quantiles by its L_p -optimal quantized version \tilde{X}^N (for some fixed N). The convergence rate of this approximation to the true conditional quantiles is obtained. Section 4 defines the corresponding estimator and proves its consistency (for the fixed- N approximated conditional quantiles). The results are illustrated on a numerical example, in which a smooth variant of the proposed estimator based on the bootstrap is also introduced. Section 5 provides some final comments. Eventually, the Appendix collects technical proofs.

2 Optimal quantization

In this section, we define the concept of L_p -norm optimal quantization and state the main results that will be used in the sequel (Section 2.1). Then we describe a stochastic algorithm that allows to perform optimal quantization (Section 2.2), and provide some convergence results for this algorithm (Section 2.3).

2.1 Definition and main results

Let X be a random d -vector defined on a probability space (Ω, \mathcal{F}, P) , with distribution P_X , and fix a real number $p \geq 1$ such that $E[|X|^p] < \infty$ (throughout, $|\cdot|$ denotes the Euclidean norm). Quantization replaces X with an appropriate random d -vector $\pi(X)$ that assumes at most N values. In optimal L_p -norm quantization, the vector $\pi(X)$ minimizes the L_p -norm quantization error

$$\|\pi(X) - X\|_p, \quad \text{with } \|Z\|_p := (E[|Z|^p])^{1/p}.$$

This optimization problem is equivalent to finding an N -grid of $\mathbb{R}^d - \gamma^N$, say — such that the projection $\tilde{X}^{\gamma^N} = \text{Proj}_{\gamma^N}(X)$ of X on the (Euclidean-)nearest point of the grid minimizes the *quantization error* $\|\tilde{X}^{\gamma^N} - X\|_p$. This definition leads to two natural questions: does such a minimum always exist? How does this minimum behave as N goes to infinity?

Existence (but not unicity) of an optimal N -grid — that is, a grid minimizing this quantization error — has been obtained under the assumption that P_X does not charge any hyperplane; see Pagès (1998). Irrespective of the sequence of optimal grids considered, \tilde{X}^N converges to X in L_p . This is a direct corollary of the following result, which is often referred to as Zador's theorem (Zador, 1964) and provides the rate of convergence of the quantization error; see, e.g., Graf and Luschgy (2000) for a proof.

Theorem 2.1. *Assume that $\|X\|_{p+\delta} < \infty$ for some $\delta > 0$. Let $P_X(du) = f(u)\lambda_d(du) + \nu(du)$ be the Lebesgue decomposition of P_X , where λ_d is the Lebesgue measure on \mathbb{R}^d and $\nu \perp \lambda_d$. Then*

$$\lim_{N \rightarrow \infty} \left(N^{\frac{p}{d}} \min_{\gamma^N \in (\mathbb{R}^d)^N} \|\tilde{X}^{\gamma^N} - X\|_p^p \right) = J_{p,d} \left(\int_{\mathbb{R}^d} (f(u))^{\frac{d}{p+d}} du \right)^{1+\frac{p}{d}},$$

with $J_{p,d} = \min_N (N^{p/d} \min_{\gamma^N \in (\mathbb{R}^d)^N} D_N^{p,U}(\gamma^N))$, where $D_N^{p,U}(\gamma^N)$ denotes the (p th power of the) quantization error, obtained for the uniform distribution over $[0, 1]^d$, when considering the grid $\gamma^N \in (\mathbb{R}^d)^N$.

In dimension $d = 1$, one has $J_{p,d} = \frac{1}{2^{p(p+1)}}$. For $d > 1$, little is known about $J_{p,d}$, but it can be shown that $J_{p,d} \sim \left(\frac{d}{2\pi e}\right)^{p/2}$ as $d \rightarrow \infty$; see Graf and Luschgy (2000).

Corollary 2.2. *Assume that $\|X\|_{p+\delta} < \infty$ for some $\delta > 0$. Then, for some $C, D \in \mathbb{R}$ and $N_0 \in \mathbb{N}$, we have that*

$$\|\tilde{X}^{\gamma^N} - X\|_p^p \leq \frac{1}{N^{p/d}} (C\|X\|_{p+\delta}^{p+\delta} + D),$$

for all $N \geq N_0$.

Summing up, there exist L_p -optimal N -grids — or *optimal N -quantizers* — that minimize the quantization error. A further natural question then is: how to obtain an optimal N -grid? We now discuss a stochastic algorithm that addresses this problem.

2.2 The stochastic gradient algorithm

Except in some very exceptional cases (such as the uniform over a compact interval of the real line), optimal N -grids have no closed form. That is, there exist no results that describe the geometric structure of such grids. However, one can attempt to obtain (approximations of) optimal N -grids through a *stochastic gradient algorithm* such as the following.

Let $(\xi_t)_{t \in \mathbb{N}_0}$, $\mathbb{N}_0 = \{1, 2, \dots\}$, be a sequence of independent and identically P_X -distributed random vectors, and let $(\delta_t)_{t \in \mathbb{N}_0}$ be a deterministic sequence in $(0, 1)$ such that

$$\sum_{t=1}^{\infty} \delta_t = +\infty \quad \text{and} \quad \sum_{t=1}^{\infty} \delta_t^2 < +\infty$$

(throughout the paper, we tacitly assume that the algorithm makes use of a sequence δ_t that satisfies these conditions). The algorithm starts from a deterministic N -tuple $X^0 = x^0$ with N pairwise distinct entries. This initial N -grid x^0 is then updated as follows. For every $t \in \mathbb{N}_0$, define recursively the grid X^t as

$$X^t = X^{t-1} - \frac{\delta_t}{p} \nabla_x d_N^p(X^{t-1}, \xi^t), \tag{2.1}$$

where $\nabla_x d_N^p(x, \xi)$ stands for the gradient with respect to the x -argument of the so-called local quantization error $d_N^p(x, \xi) = \min_{1 \leq i \leq N} |x_i - \xi|^p$, with $x = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N$ and $\xi \in \mathbb{R}^d$. Note that, for any ξ , the i th entry of this gradient is given by

$$(\nabla_x d_N^p(x, \xi))_i = p |x_i - \xi|^{p-1} \frac{x_i - \xi}{|x_i - \xi|} \mathbb{I}_{[x_i = \text{Proj}_x(\xi)]},$$

where \mathbb{I}_A denotes the indicator function of the set A , and with the convention $0/0 = 1$ when $x_i = \xi$. This implies that the N -vector $\nabla_x d_N^p(x, \xi)$ always has exactly one non-zero entry, namely the one corresponding to the point of the grid x that is closest to ξ . Consequently, at each step t of the algorithm, only one point of the grid X^{t-1} will be changed to define the grid X^t , namely the point from the grid X^{t-1} that is closest to ξ^t .

More details about this algorithm can be found in [Pagès and Printems \(2003\)](#). For $p = 2$, this is known as the *Competitive Learning Vector Quantization (CLVQ)* algorithm, and is the most commonly used one in quantization. This success is explained by the fact that the convergence results obtained for the CLVQ algorithm are much more satisfactory than for $p \neq 2$.

2.3 Convergence results for the CLVQ algorithm

Here we state several results showing that the grids provided by the CLVQ algorithm converge to optimal grids as the number of iterations t goes to infinity.

We start with the univariate case ($d = 1$). Assume that the support of P_X is compact and let its convex hull C be $[a, b]$. Write $F_N^+ := \{x = (x_1, \dots, x_N) : a < x_1 < \dots < x_N < b\}$ for the set of N -grids on C involving pairwise distinct points stored in ascending order, and let \bar{F}_N^+ be its closure; see [Pagès \(1998\)](#). Denote by $D_N^{2, P_X}(x) = \int_C \min_{1 \leq i \leq N} |x_i - w|^2 P_X(dw)$ the (squared) L_2 -norm quantization error associated with a given grid $x = (x_1, \dots, x_N) \in F_N^+$.

Theorem 2.3 ([Pagès, 1998](#), Th. 27). *In the univariate setup above, we have the following.*

- (i) *Assume that P_X is absolutely continuous with a density $f : [a, b] \rightarrow \mathbb{R}^+$ that is positive on (a, b) , and assume either that f is strictly log-concave or that it is log-concave with $f(a+) - f(b-) > 0$. Then $x \mapsto D_N^{2, P_X}(x)$ has a unique minimizer x^* in \bar{F}_N^+ , which coincides with the unique solution of $\nabla D_N^{2, P_X}(x) = 0$ in \bar{F}_N^+ (when P_X is the uniform over $[0, 1]$, the optimal grid is $x^* = (a + \frac{2k-1}{2N}(b-a))_{1 \leq k \leq N}$).*
- (ii) *Irrespective of the initial grid $X^0 \in F_N^+$, every trajectory (X^0, X^1, X^2, \dots) of the CLVQ algorithm is a.s. such that $X^t \in F_N^+$ for all t . If P_X is absolutely continuous and if there are finitely many grids $x \in \bar{F}_N^+$ such that $\nabla D_N^{2, P_X}(x) = 0$, then $X^t \xrightarrow{a.s.} x^*$ as $t \rightarrow \infty$, with $\nabla D_N^{2, P_X}(x^*) = 0$.*

Part (i) of the result provides a particular family of distributions for which the optimal grid is unique (recall that existence always holds). Beyond stating that trajectories of the CLVQ algorithm live in F_N^+ (with grids that therefore stays of size N), Part (ii) of the result provides mild conditions under which the algorithm almost surely provides a limiting grid that is a critical point of the quantization error, hence, under the assumptions of Part (i), is optimal.

Unfortunately, the picture is less clear for the multivariate case ($d > 1$). While it is still so that the grid X^t will have pairwise distinct components for any t , some of the components of the limiting grid x^* , if any, may coincide.

- (a) If, parallel to the univariate case, this does not happen, then the a.s. convergence of X^t to a critical point of the quantization error $D_N^{2, P_X}(\cdot)$ can be established under the assumption

that P_X has a bounded density with a compact and convex support.

- (b) Otherwise, no convergence results are available; the only optimality results that can then be obtained relate to approximations involving grids of size $k < N$, where k is the number of distinct components in the limiting grid x^* , which is quite different from the original N -quantization problem considered initially.

The interested reader may refer to [Pagès \(1998\)](#) for details. For practical purposes, though, one should not worry too much, as all numerical exercises we conducted were compatible with case (a) (with increasing t , the smallest distance between two components of X^t always seemed to stabilize rather than decreasing to zero).

3 Conditional quantiles through optimal quantization

Let us come back to the regression setup involving a scalar response Y and a d -dimensional vector of covariates X , and consider the conditional quantile functions $q_\alpha(\cdot)$ in (1.2). It is well-known that

$$q_\alpha(x) = \arg \min_{a \in \mathbb{R}} \mathbb{E}[\rho_\alpha(Y - a) | X = x], \quad (3.1)$$

where $z \mapsto \rho_\alpha(z) = -(1 - \alpha)z\mathbb{I}_{[z < 0]} + \alpha z\mathbb{I}_{[z \geq 0]} = z(\alpha - \mathbb{I}_{[z < 0]})$ is the so-called *check function*. As we now explain, this allows to use optimal quantization to approximate conditional quantiles.

To do so, fix $p \geq 1$ such that $\|X\|_p < \infty$. Then, for any positive integer N , one may consider the approximation

$$\tilde{q}_\alpha^N(x) = \arg \min_{a \in \mathbb{R}} \mathbb{E}[\rho_\alpha(Y - a) | \tilde{X}^N = \tilde{x}], \quad (3.2)$$

where \tilde{X}^N and \tilde{x} are the projections of X and x respectively onto an L_p -optimal N -grid. Since $\tilde{X}^N - X$ goes to zero as $N \rightarrow \infty$, one may expect that $\tilde{q}_\alpha^N(x)$ provides a better and better approximation of $q_\alpha(x)$ as N increases. The main goal of this section is to quantify the quality of this approximation.

We will need the following assumptions.

ASSUMPTION (A) (i) The random vector (X, Y) is generated through $Y = m(X, \varepsilon)$, where the d -dimensional covariate vector X and the error ε are mutually independent; (ii) the link function $(x, z) \mapsto m(x, z)$ is of the form $m_1(x) + m_2(x)z$, where the functions $m_1(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ and $m_2(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_0^+$ are Lipschitz functions; (iii) $\|X\|_p < \infty$ and $\|\varepsilon\|_p < \infty$; (iv) the distribution of X does not charge any hyperplane.

Note that Assumption (A)(ii)-(iii) directly implies that there exists $C > 0$ such that the link

function $m(\cdot, \varepsilon)$ of the model above satisfies

$$\forall u, v \in \mathbb{R}^d, \quad \|m(u, \varepsilon) - m(v, \varepsilon)\|_p \leq C|u - v|. \quad (3.3)$$

The resulting Lipschitz constant — that is, the smallest real number C for which (3.3) holds — is $[m]_{\text{Lip}} = [m_1]_{\text{Lip}} + [m_2]_{\text{Lip}}\|\varepsilon\|_p$, where $[m_1]_{\text{Lip}}$ and $[m_2]_{\text{Lip}}$ are the corresponding Lipschitz constants of m_1 and m_2 , respectively.

ASSUMPTION (B) (i) The support S_X of P_X is compact; (ii) ε admits a continuous density $f^\varepsilon : \mathbb{R} \rightarrow \mathbb{R}_0^+$ (with respect to the Lebesgue measure on \mathbb{R}).

To obtain rates of convergence, we will need the following reinforcement of Assumption (A).

ASSUMPTION (A') Same as Assumption (A), but with (iii) replaced by (iii)' there exists $\delta > 0$ such that $\|X\|_{p+\delta} < \infty$, and $\|\varepsilon\|_p < \infty$.

We can then prove the following result (see the Appendix for the proof).

Theorem 3.1. *Fix $\alpha \in (0, 1)$. Then (i) under Assumptions (A)-(B),*

$$\|\tilde{q}_\alpha^N(X) - q_\alpha(X)\|_p \leq 2 \sqrt{\max\left(\frac{\alpha}{1-\alpha}, \frac{1-\alpha}{\alpha}\right)} [m]_{\text{Lip}}^{1/2} \|L^N(X)\|_p^{1/2} \|X - \tilde{X}^N\|_p^{1/2},$$

for N sufficiently large, where $(L^N(X))$ is a sequence of X -measurable random variables that is bounded in L_p ; (ii) under Assumptions (A')-(B),

$$\|\tilde{q}_\alpha^N(X) - q_\alpha(X)\|_p = O(N^{-1/2d}),$$

as $N \rightarrow \infty$.

Of course, fixed- x consistency results are also quite appealing in quantile regression. Such a result is provided in the following theorem (see the Appendix for the proof).

Theorem 3.2. *Fix $\alpha \in (0, 1)$. Then, under Assumptions (A)-(B),*

$$\sup_{x \in S_X} |\tilde{q}_\alpha^N(x) - q_\alpha(x)| \rightarrow 0,$$

as $N \rightarrow \infty$.

Unlike in Theorem 3.1, Theorem 3.2 does not provide any rate of convergence. This is a consequence of the fact that, while the convergence of \tilde{X}^N towards X can be shown to imply the convergence of \tilde{x} towards x for each fixed x , it does not seem possible to show that the rate of convergence in the fixed- x convergence is inherited from the convergence involving X .

4 Quantized conditional quantile estimators

4.1 The proposed estimators and their consistency

Consider now the problem of estimating the conditional quantile $q_\alpha(x)$ on the basis of independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of (X, Y) . For any $N(< n)$, the approximation $\tilde{q}_\alpha^N(x)$ in (3.2) leads to an estimator $\hat{q}_\alpha^{N,n}(x)$ of the conditional quantile $q_\alpha(x)$, through the following two steps :

(S1) First, the CLVQ algorithm from Section 2.2 is applied to perform quantization in X . For this purpose, (i) the initial grid X^0 is obtained by sampling randomly among the X_i 's without replacement, and with the constraint¹ that the same x -values cannot be picked more than once; (ii) n iterations are performed, based on $\xi^t = X_t$, $t = 1, \dots, n$. We write $\hat{\gamma}^{N,n} = (\hat{x}_1^{N,n}, \dots, \hat{x}_N^{N,n})$ for the resulting grid and $\hat{X}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(X)$ for the corresponding (empirical) quantization of X ; to make the notation less heavy, we will stress dependence on n in these quantities only when it is necessary.

(S2) Second, the approximation $\tilde{q}_\alpha^N(x) = \arg \min_a E[\rho_\alpha(Y - a) | \tilde{X}^N = \tilde{x}]$ is then estimated by

$$\hat{q}_\alpha^{N,n}(x) = \arg \min_a \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{I}_{[\hat{X}_i^N = \hat{x}^N]},$$

where $\hat{X}_i^N = \hat{X}_i^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(X_i)$ and $\hat{x}^N = \hat{x}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(x)$. Of course, $\hat{q}_\alpha^{N,n}(x)$, in practice, is simply evaluated as the sample α -quantile of the Y_i 's whose corresponding \hat{X}_i^N is equal to \hat{x}^N .

Note that the number of iterations is equal to the sample size n at hand, so that it is expected that only moderate-to-large n will provide reasonable approximations of optimal N -grids.

For fixed N (and x), the convergence in probability of $\hat{q}_\alpha^{N,n}(x)$ to $\tilde{q}_\alpha^N(x)$ as $n \rightarrow \infty$ can be obtained by making use of the convergence results for the stochastic gradient algorithm discussed in Section 2. In order to do so, we need to restrict to $p = 2$ (that is, to the CLVQ algorithm) and to adopt the following assumption.

ASSUMPTION (C) P_X is absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^d .

We then have the following result.

Theorem 4.1. *Fix $\alpha \in (0, 1)$, $x \in S_X$ and $N \in \mathbb{N}_0$. Then, under Assumptions (A), (B)(i), and (C), we have that, as $n \rightarrow \infty$,*

$$|\hat{q}_\alpha^{N,n}(x) - \tilde{q}_\alpha^N(x)| \rightarrow 0,$$

¹If the X_i 's are i.i.d. with a common density f , sampling without replacement among the X_i 's of course implies that this constraint will be met with probability one. One often needs to impose it, however, in real-data examples (due to the possible presence of ties) or when performing bootstrap (see later).

in probability, provided that quantization is based on $p = 2$.

In the previous section, we showed that, as $N \rightarrow \infty$, $\tilde{q}_\alpha^N(x) - q_\alpha(x)$ goes to zero almost surely, hence in probability. Theorem 4.1 then suggests that $\hat{q}_\alpha^{N,n}(x) - q_\alpha(x)$ might go to zero in probability as both n and N go to infinity in some appropriate way. Obtaining such a double asymptotic results, however, is extremely delicate, since, to the best of our knowledge, all convergence results for the stochastic gradient algorithm in the literature are as $n \rightarrow \infty$ with N fixed.

4.2 Numerical example and bootstrap modification

For the sake of illustration, we evaluated the estimator $\hat{q}_\alpha^{N,n}(x)$ for $N = 10, 25$, and 40 , in a sample of $n = 300$ mutually independent observations (X_i, Y_i) obtained from the model

$$Y = \frac{1}{5}X^3 + \varepsilon, \quad (4.1)$$

where $X = 6Z - 3$, with $Z \sim \text{Beta}(0.3, 0.3)$, and $\varepsilon \sim \mathcal{N}(0, 1)$ are independent. The left panels in Figure 1 plot the corresponding quantiles curves $x \mapsto \hat{q}_\alpha^{N,n}(x)$ for $\alpha = 0.05, 0.25, 0.5, 0.75$, and 0.95 (actually, these curves were evaluated only at 300 equispaced points in $(-3, 3)$). It is seen that the number of quantizers N used has an important impact on the curves. For small N , the curves are not smooth and show a large bias. For large N , bias is smaller but the variability is large. One should keep in mind that, for large N , the grid provided by the CLVQ algorithm poorly approximates the corresponding optimal N -grid, since a fixed number n of iterations are used in the algorithm (that should not be too small compared to N).

Smoother quantile curves can be obtained from the bootstrap, through the following procedure. For some integer B , we first generate grids $\hat{\gamma}_b^{N,n}$, $b = 1, \dots, B$ (each of size N), as follows from the CLVQ algorithm: first, we sample N observations with replacement from the initial sample X_1, \dots, X_n to generate an initial grid X_b^0 (with the same constraint as in (S1) above that the N values obtained are pairwise different); second, we perform iterations based on ξ_b^t , $t = 1, \dots, n$, that are similarly obtained by sampling with replacement from X_1, \dots, X_n (this time, without any constraint). This allows to consider the bootstrap estimators

$$\bar{q}_{\alpha,B}^{N,n}(x) = \frac{1}{B} \sum_{b=1}^B \hat{q}_\alpha^{(b)}(x),$$

where $\hat{q}_\alpha^{(b)}(x) = \hat{q}_\alpha^{(b),N,n}(x)$ is obtained by performing (S2) based on the original sample (X_i, Y_i) , $i = 1, \dots, n$, and the grid $\hat{\gamma}_b^{N,n}$. Bootstrapping, thus, focuses on the construction of the grids.

The right panels of Figure 1 plot the resulting bootstrapped quantile curves $x \mapsto \bar{q}_{\alpha,B}^{N,n}(x)$ for the same values of α and N as in the original (non-bootstrapped) versions. Bootstrapping

clearly smooths all curves, and moreover significantly reduces boundary effects for small N . These advantages require to take B large enough. But of course, very large values of B should be avoided in order to keep the computational burden under control.

The main competitors to the proposed quantization-based estimators are the nearest-neighbor estimators (as those of [Bhattacharya and Gangopadhyay, 1990](#)) and the local constant and local linear estimators from [Yu and Jones \(1998\)](#). For the sake of comparison, we plot those competitors in Figure 2, jointly with our bootstrapped estimator (based on $B = 50$), on the same sample used in Figure 1. For each estimator, the smoothing parameters involved were selected in an automatic way. For nearest-neighbor estimators, the number k of neighbors to consider was chosen to minimize the MSE

$$\text{MSE}(k) = \frac{1}{N_x} \sum_{i=1}^{N_x} (\hat{q}_\alpha^k(x_i) - q_\alpha(x_i))^2,$$

where $\hat{q}_\alpha^k(x)$ denotes the nearest-neighbor estima of $q_\alpha(x)$ based on k neighbors and $\{x_1, \dots, x_{N_x}\}$ is an equispaced grid on $(-3, 3)$ (note that this method uses the population conditional quantiles, hence is infeasible in practice). Local constant/linear estimators (that are based on the Gaussian kernel) involve a bandwidth that, for fixed α , is selected (here in a genuinely data-driven way) as

$$h_\alpha = \frac{\alpha(1-\alpha)h_{\text{mean}}}{\phi(\Phi^{-1}(\alpha))^2},$$

where ϕ and Φ are respectively the standard normal density and distribution functions. Here, h_{mean} corresponds to the optimal bandwidth for regression mean estimation and is chosen through cross-validation; see [Yu and Jones \(1998\)](#). As for the quantization-based estimators, the number of quantizers N is selected through a data-driven method that is defined and investigated in the companion paper [Charlier et al. \(2014\)](#).

Since Z is generated according to a beta distribution with (equal) parameter values that are smaller than one, the X_i 's are less dense in the middle of the interval $(-3, 3)$ than close to its boundaries. The local constant and linear estimators seem to suffer from this fact, as the corresponding quantiles curves are significantly less smooth in the middle than at the boundaries. The nearest-neighbor estimator and our quantization-based estimator provide better estimations in the middle part, with an advantage for our estimator that appears to be smoother (despite the fact that, unlike for the nearest-neighbor estimator, the corresponding smoothing parameter value is chosen in a totally data-driven way).

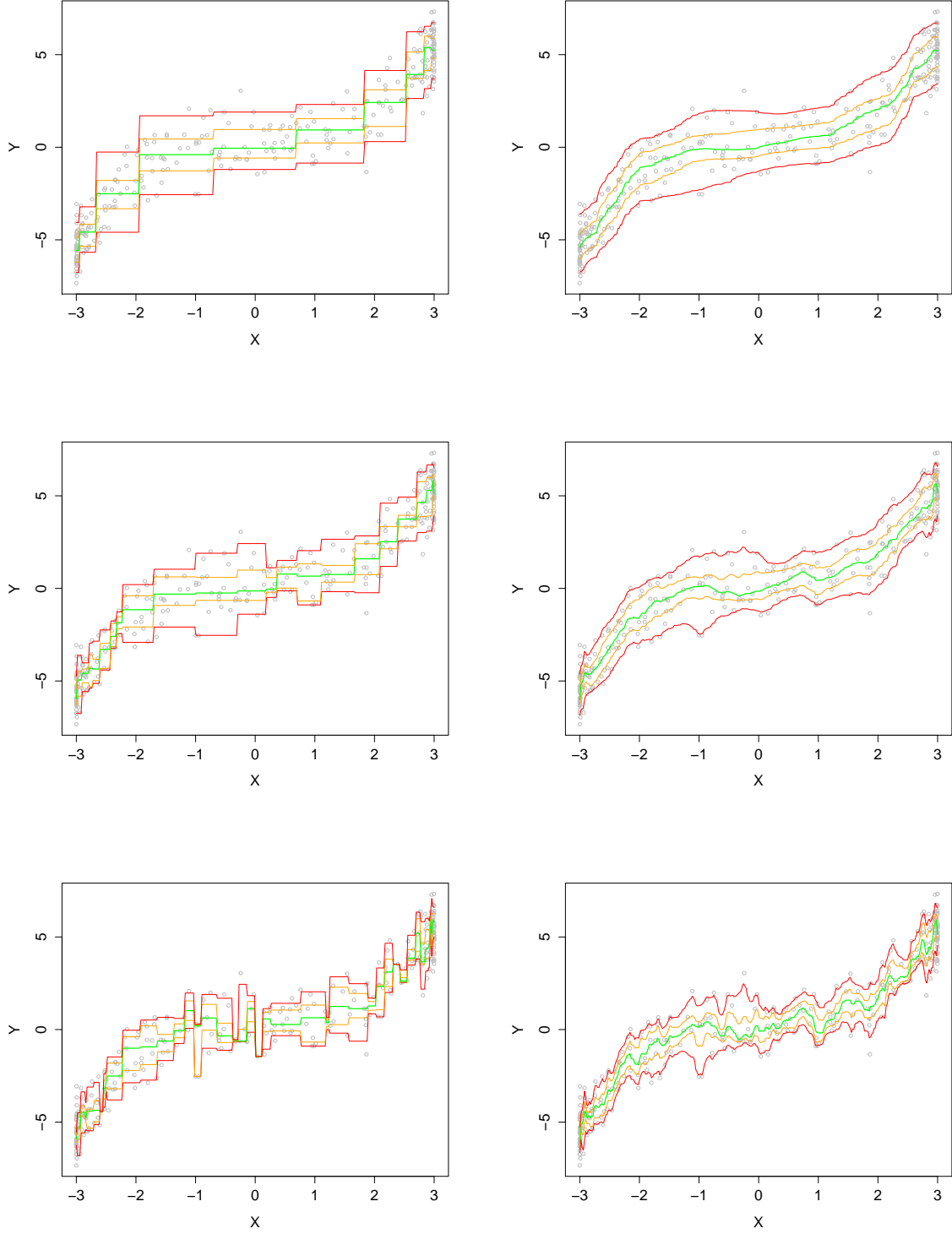


Figure 1: Estimated conditional quantiles curves $x \mapsto \hat{q}_\alpha^{N,n}(x)$ (left) and their bootstrapped counterparts $x \mapsto \hat{q}_{\alpha,B}^{N,n}(x)$ for $B = 50$ (right), based on $N = 10$ (top), $N = 25$ (middle), and $N = 50$ (bottom). The sample size is $n = 300$, and the quantiles levels considered are $\alpha=0.05, 0.25, 0.5, 0.75$, and 0.95 . See (4.1) for the data generating model.

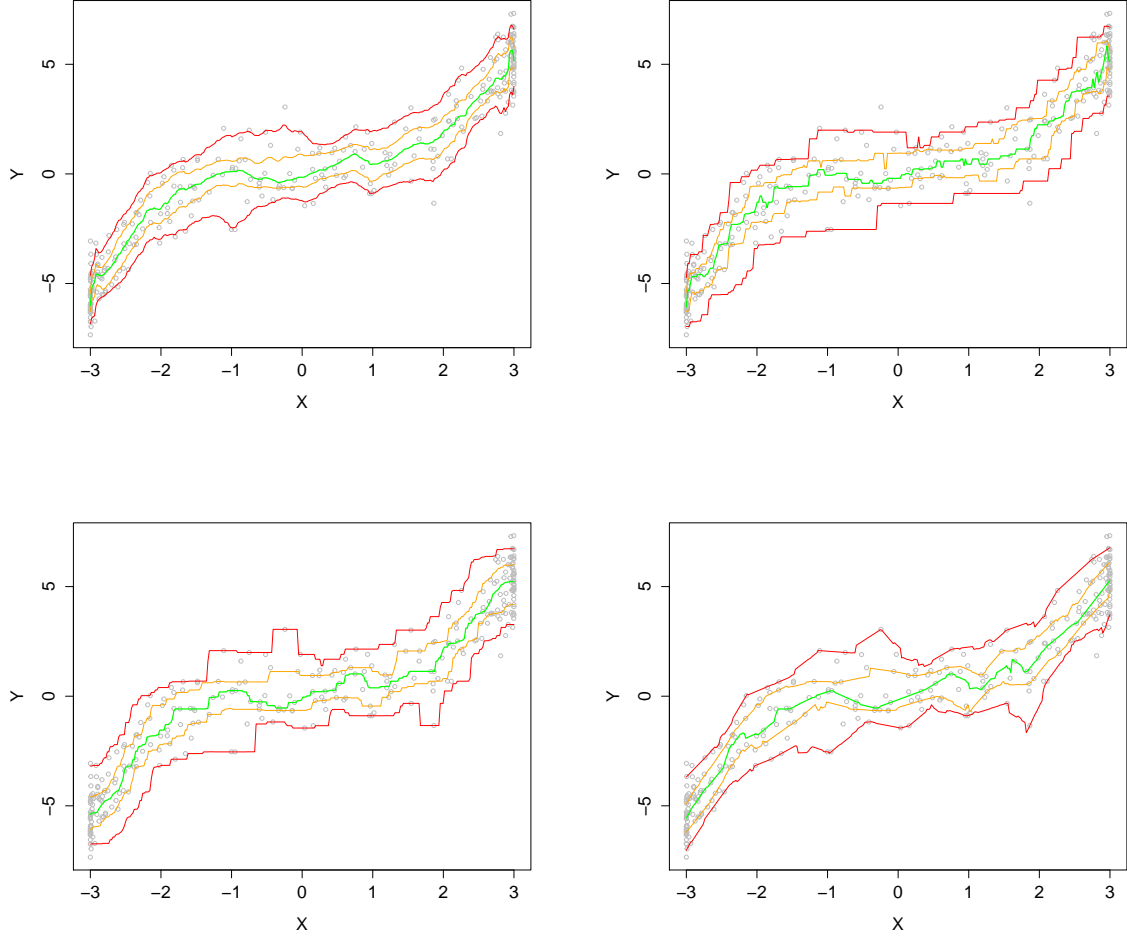


Figure 2: The proposed estimated conditional quantiles curves $x \mapsto \hat{q}_{\alpha, B}^{N, n}(x)$ for $B = 50$ (top left), their nearest-neighbor competitors from [Bhattacharya and Gangopadhyay \(1990\)](#) (top right), and their local constant and local linear competitors from [Yu and Jones \(1998\)](#) (bottom left and bottom right, respectively). The sample is the same as in Figure 1.

5 Final comments

In this paper, we presented a new method to estimate nonparametrically conditional quantile curves of Y given X . The main idea is to use optimal quantization as an alternative to more standard localization techniques such as kernel or nearest-neighbor methods. Our construction consists in replacing X in the definition of conditional quantiles with a quantized version of X . We showed that this provides a valid approximation of conditional quantiles. In the empirical case, this approximation leads to new estimators of conditional quantiles, that, as we proved, are (for fixed N) consistent to their population counterparts.

We illustrated the behavior of these estimators on a numerical example, and showed that N essentially behaves as a smoothing parameter, parallel to the bandwidth and the number of neighbors to be used for kernel and nearest-neighbor methods, respectively. Of course, extensive simulations are needed to compare the performances of the proposed estimators with these classical competitors. This is achieved in the companion paper [Charlier et al. \(2014\)](#), where it is shown that quantization-based estimators tend to dominate their competitors in terms of MSEs as soon as X is not uniformly distributed. There, a method to choose empirically the smoothing parameter N is also developed and investigated.

A Proofs of Section 3

In this section, we prove Theorems 3.1 and 3.2, which requires to establish several lemmas. We first introduce some notation. Let $G_a(x) = \mathbb{E}[\rho_\alpha(Y - a)|X = x]$ and denote the corresponding quantized quantity by $\tilde{G}_a(\tilde{x}) = \mathbb{E}[\rho_\alpha(Y - a)|\tilde{X}^N = \tilde{x}]$. Since $\tilde{q} = \tilde{q}_\alpha^N(x) = \arg \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})$ and $q = q_\alpha(x) = \arg \min_{a \in \mathbb{R}} G_a(x)$, it is natural to try to control the distance between $\tilde{G}_a(\tilde{x})$ and $G_a(x)$. This is achieved in Lemma A.2, the proof of which requires the following preliminary result.

Lemma A.1. *Fix $\alpha \in (0, 1)$ and $a \in \mathbb{R}$. Then, under Assumption (A), (i) $\rho_\alpha : \mathbb{R} \rightarrow \mathbb{R}$ is Lipschitz, with Lipschitz constant $[\rho_\alpha]_{\text{Lip}} = \max(\alpha, 1 - \alpha)$, and (ii) $G_a : \mathbb{R}^d \rightarrow \mathbb{R}$ is Lipschitz, with Lipschitz constant $[G_a]_{\text{Lip}} = \max(\alpha, 1 - \alpha)[m]_{\text{Lip}}$.*

Proof of Lemma A.1. Since (i) is a trivial calculus exercise, we only prove (ii). To do so, note that, for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned} |G_a(u) - G_a(v)| &= |\mathbb{E}[\rho_\alpha(Y - a)|X = u] - \mathbb{E}[\rho_\alpha(Y - a)|X = v]| \\ &= |\mathbb{E}[\rho_\alpha(m(X, \varepsilon) - a)|X = u] - \mathbb{E}[\rho_\alpha(m(X, \varepsilon) - a)|X = v]|, \end{aligned}$$

so that the independence of X and ε entails

$$|G_a(u) - G_a(v)| \leq \mathbb{E}[|\rho_\alpha(m(u, \varepsilon) - a) - \rho_\alpha(m(v, \varepsilon) - a)|] \leq [\rho_\alpha]_{\text{Lip}}[m]_{\text{Lip}}|u - v|,$$

where the last inequality follows from Part (i) of the result and equation (3.3). \square

We still need the following lemma to prove Theorem 3.2.

Lemma A.2. *Fix $\alpha \in (0, 1)$ and $x \in S_X$. For any integer N , let $\tilde{x} = \tilde{x}^N = \text{Proj}_{\gamma^N}(x)$ and $C_x = C_x^N = \{z \in S_X : \text{Proj}_{\gamma^N}(z) = \tilde{x}\}$. Then, under Assumptions (A)-(B),*

- (i) $\sup_{x \in S_X} |x - \tilde{x}| \rightarrow 0$ as $N \rightarrow \infty$;
- (ii) $\sup_{x \in S_X} R(C_x) \rightarrow 0$ as $N \rightarrow \infty$, where we let $R(C_x) := \sup_{z \in C_x} |z - \tilde{x}|$;
- (iii) $\sup_{x \in S_X} \sup_{a \in \mathbb{R}} |\tilde{G}_a(\tilde{x}) - G_a(x)| \rightarrow 0$ as $N \rightarrow \infty$;
- (iv) $\sup_{x \in S_X} |\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)| \rightarrow 0$ as $N \rightarrow \infty$.

Proof of Lemma A.2. (i) Assume by contradiction that there exists $\varepsilon > 0$ such that, for infinitely many N (for $N \in \mathcal{N}(\varepsilon)$, say), we have $\sup_{x \in S_X} |\tilde{x}^N - x| > \varepsilon$. For any such value of N , one can pick $x \in S_X$ (that may depend on N) with $|\tilde{x}^N - x| > \varepsilon$. No point of the optimal grid γ^N belongs to the ball $\mathcal{B}(x, \varepsilon) = \{z \in \mathbb{R}^d : |z - x| < \varepsilon\}$, which implies that, for all $z \in \mathcal{B}(x, \varepsilon/2)$, $|\tilde{z}^N - z| > \varepsilon/2$, where \tilde{z} is the projection of z onto γ^N . Therefore, for $N \in \mathcal{N}(\varepsilon)$,

$$\begin{aligned} \|X^{\gamma^N} - X\|_p^p &= \int_{\text{Supp}(P_X)} |\tilde{z}^N - z|^p dP_X(z) \geq \\ &\int_{\mathcal{B}(x, \varepsilon/2)} |\tilde{z}^N - z|^p dP_X(z) > \left(\frac{\varepsilon}{2}\right)^p \inf_{y \in S_X} P_X[\mathcal{B}(y, \varepsilon/2)] =: \delta_\varepsilon > 0, \end{aligned} \quad (\text{A.1})$$

where the last inequality follows from the fact that $y \mapsto P_X[\mathcal{B}(y, \varepsilon/2)]$ is a continuous function taking only strictly positive values on the compact set S_X . Since the cardinality of $\mathcal{N}(\varepsilon)$ is infinite, (A.1) prevents $\|X^{\gamma^N} - X\|_p$ to go to zero as $N \rightarrow \infty$, a contradiction.

(ii) Since S_X is compact, we have that, for any x , the radius of the quantization cell C_x is bounded. Hence, for any x , there exists $x^{*N} \in C_x$ such that

$$R(C_x) = \sup_{z \in C_x} |z - \tilde{x}| = |x^{*N} - \tilde{x}| \leq \sup_{x \in S_X} |x - \tilde{x}|.$$

Hence,

$$\sup_{x \in S_X} R(C_x) \leq \sup_{x \in S_X} |x - \tilde{x}|.$$

The result then follows from Part (i).

(iii) Fix $a \in \mathbb{R}$. First note that, since $[\tilde{X}^N = \tilde{x}]$ is equivalent to $[X \in C_x]$, one has

$$|\mathbb{E}[\rho_\alpha(Y - a)|\tilde{X}^N = \tilde{x}] - \mathbb{E}[\rho_\alpha(Y - a)|X = \tilde{x}]| \leq \sup_{z \in C_x} |\mathbb{E}[\rho_\alpha(Y - a)|X = z] - \mathbb{E}[\rho_\alpha(Y - a)|X = \tilde{x}]|.$$

Therefore,

$$\begin{aligned}
& |\tilde{G}_a(\tilde{x}) - G_a(x)| = |\mathbb{E}[\rho_\alpha(Y - a)|\tilde{X}^N = \tilde{x}] - \mathbb{E}[\rho_\alpha(Y - a)|X = x]| \\
& \leq |\mathbb{E}[\rho_\alpha(Y - a)|\tilde{X}^N = \tilde{x}] - \mathbb{E}[\rho_\alpha(Y - a)|X = \tilde{x}]| + |\mathbb{E}[\rho_\alpha(Y - a)|X = \tilde{x}] - \mathbb{E}[\rho_\alpha(Y - a)|X = x]| \\
& \leq 2 \sup_{z \in C_x} |\mathbb{E}[\rho_\alpha(Y - a)|X = z] - \mathbb{E}[\rho_\alpha(Y - a)|X = \tilde{x}]|.
\end{aligned}$$

Using the independence between X and ε and the Lipschitz properties of ρ_α and m then yields

$$\begin{aligned}
|\tilde{G}_a(\tilde{x}) - G_a(x)| & \leq 2 \sup_{z \in C_x} |\mathbb{E}[\rho_\alpha(m(z, \varepsilon) - a) - \mathbb{E}[\rho_\alpha(m(\tilde{x}, \varepsilon) - a)]| \\
& \leq 2 \max(\alpha, 1 - \alpha)[m]_{\text{Lip}} \sup_{z \in C_x} |z - \tilde{x}| = 2 \max(\alpha, 1 - \alpha)[m]_{\text{Lip}} R(C_x). \tag{A.2}
\end{aligned}$$

Hence,

$$\sup_{x \in S_X} \sup_{a \in \mathbb{R}} |\tilde{G}_a(\tilde{x}) - G_a(x)| \leq 2 \max(\alpha, 1 - \alpha)[m]_{\text{Lip}} \sup_{x \in S_X} R(C_x).$$

The results then follows from Part (ii).

(iv) Letting $\mathbb{I}_+ = \mathbb{I}_{[\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) \geq \min_{a \in \mathbb{R}} G_a(x)]}$, we have

$$\begin{aligned}
& |\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)| \mathbb{I}_+ = (\tilde{G}_{\tilde{q}_\alpha^N(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)) \mathbb{I}_+ \\
& \leq (\tilde{G}_{q_\alpha(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)) \mathbb{I}_+ \leq \sup_{a \in \mathbb{R}} |\tilde{G}_a(\tilde{x}) - G_a(x)| \mathbb{I}_+.
\end{aligned}$$

Proceeding similarly with $\mathbb{I}_- = \mathbb{I}_{[\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) < \min_{a \in \mathbb{R}} G_a(x)]}$, this yields

$$\begin{aligned}
& |\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)| \mathbb{I}_- = (G_{q_\alpha(x)}(x) - \tilde{G}_{\tilde{q}_\alpha^N(x)}(\tilde{x})) \mathbb{I}_- \\
& \leq (G_{\tilde{q}_\alpha^N(x)}(x) - \tilde{G}_{\tilde{q}_\alpha^N(x)}(\tilde{x})) \mathbb{I}_- \leq \sup_{a \in \mathbb{R}} |\tilde{G}_a(\tilde{x}) - G_a(x)| \mathbb{I}_-,
\end{aligned}$$

so that $|\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)| \leq \sup_{a \in \mathbb{R}} |\tilde{G}_a(\tilde{x}) - G_a(x)|$. The result therefore follows from Part (iii). \square

We can now prove Theorem 3.2.

Proof of Theorem 3.2. First note that, for any $x \in S_X$,

$$\begin{aligned}
& |G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \leq |G_{\tilde{q}_\alpha^N(x)}(x) - \tilde{G}_{\tilde{q}_\alpha^N(x)}(\tilde{x})| + |\tilde{G}_{\tilde{q}_\alpha^N(x)}(\tilde{x}) - G_{q_\alpha(x)}(x)| \\
& \leq \sup_{a \in \mathbb{R}} |G_a(x) - \tilde{G}_a(\tilde{x})| + |\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)| \\
& \leq \sup_{x \in S_X} \sup_{a \in \mathbb{R}} |G_a(x) - \tilde{G}_a(\tilde{x})| + \sup_{x \in S_X} |\min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) - \min_{a \in \mathbb{R}} G_a(x)|.
\end{aligned}$$

Therefore, Lemma A.2(iii)-(iv) readily implies that, as $N \rightarrow \infty$,

$$\sup_{x \in S_X} |G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \rightarrow 0. \quad (\text{A.3})$$

Now, let \tilde{N} be such that, for any $N \geq \tilde{N}$, we have $|G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \leq 1$ for all $x \in S_X$. As we will show later in this proof, this implies that there exists M such that

$$|\tilde{q}_\alpha^N(x) - q_\alpha(x)| \leq M, \quad (\text{A.4})$$

for all $x \in S_X$ and $N \geq \tilde{N}$.

One can easily check that, for any $x \in S_X$, $a \mapsto G_a(x)$ is twice continuously differentiable, with derivatives

$$\frac{dG_a(x)}{da} = F^\varepsilon\left(\frac{a - m_1(x)}{m_2(x)}\right) - \alpha \quad \text{and} \quad \frac{d^2G_a(x)}{da^2} = \frac{1}{m_2(x)} f^\varepsilon\left(\frac{a - m_1(x)}{m_2(x)}\right).$$

Consequently, performing a second-order expansion about $a = q_\alpha(x)$ provides

$$G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x) = \frac{1}{m_2(x)} f^\varepsilon\left(\frac{q_{\alpha*}^N(x) - m_1(x)}{m_2(x)}\right) (\tilde{q}_\alpha^N(x) - q_\alpha(x))^2,$$

for some $q_{\alpha*}^N(x)$ between $\tilde{q}_\alpha^N(x)$ and $q_\alpha(x)$. Therefore,

$$\sup_{x \in S_X} (\tilde{q}_\alpha^N(x) - q_\alpha(x))^2 \leq \frac{\sup_{x \in S_X} m_2(x)}{\inf_{x \in S_X} f^\varepsilon\left(\frac{q_{\alpha*}^N(x) - m_1(x)}{m_2(x)}\right)} \sup_{x \in S_X} |G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)|. \quad (\text{A.5})$$

Since $m_2(\cdot)$ is a continuous function defined over the compact set S_X , we have

$$\sup_{x \in S_X} m_2(x) \leq C_a \quad (\text{A.6})$$

for some constant C_a . Using (A.4) and the fact that $q_\alpha(\cdot)$, $m_1(\cdot)$, and $m_2(\cdot)$ are continuous functions also defined over this compact set (with $m_2(\cdot)$ taking strictly positive values), we have that, for $N \geq \tilde{N}$,

$$\sup_{x \in S_X} \frac{|q_{\alpha*}^N(x) - m_1(x)|}{m_2(x)} \leq \frac{\sup_{x \in S_X} (|q_\alpha(x)| + |m_1(x)| + M)}{\inf_{x \in S_X} m_2(x)} \leq C_b,$$

for some constant C_b . Jointly with the continuity of the positive function $f^\varepsilon(\cdot)$, this implies that the infimum in (A.5) admits a strictly positive lower bound that is independent of N (for $N \geq \tilde{N}$). Using this, (A.3) and (A.6), we conclude from (A.5) that

$$\sup_{x \in S_X} (\tilde{q}_\alpha^N(x) - q_\alpha(x))^2 \rightarrow 0,$$

as $N \rightarrow \infty$, which was to be proved.

It remains to show that the claim in (A.4) indeed holds true. By contradiction, assume that for all M , there exists $x = x_M$ (and $N \geq \tilde{N}$) such that $|\tilde{q}_\alpha^N(x) - q_\alpha(x)| > M$. The convexity

of $a \mapsto G_a(x)$ and the fact that $|G_{\tilde{q}_\alpha^N(x)}(x) - G_{q_\alpha(x)}(x)| \leq 1$ for all $x \in S_X$ (for $N \geq \tilde{N}$) readily implies that, for any a with $|a - q_\alpha(x)| \leq M$,

$$G_a(x) \leq G_{q_\alpha(x)}(x) + \frac{|a - q_\alpha(x)|}{M}.$$

Since this holds for all arbitrarily large M , the convexity of $G_a(x)$ implies that

$$f^\varepsilon \left(\frac{q_\alpha(x) + 1 - m_1(x)}{m_2(x)} \right) = \left. \frac{d^2 G_a(x)}{da^2} \right|_{a=q_\alpha(x)+1} \rightarrow 0$$

as $M \rightarrow \infty$. The same argument as above, however, shows that $\inf_{x \in S_X} f^\varepsilon \left(\frac{q_\alpha(x) + 1 - m_1(x)}{m_2(x)} \right) > 0$, a contradiction. \square

The proof of Theorem 3.1 requires the following three lemmas.

Lemma A.3. *Fix $\alpha \in (0, 1)$. Then (i) under Assumptions (A)-(B), we have*

$$\left\| \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)| \right\|_p \leq 2 \max(\alpha, 1 - \alpha) [m]_{\text{Lip}} \|\tilde{X}^N - X\|_p; \quad (\text{A.7})$$

(ii) under Assumptions (A')-(B), we have that

$$\left\| \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)| \right\|_p = O(N^{-1/d}),$$

as $N \rightarrow \infty$.

Lemma A.4. *Fix $\alpha \in (0, 1)$. Then (i) under Assumptions (A)-(B), we have $\|\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)\|_p \leq 2 \max(\alpha, 1 - \alpha) [m]_{\text{Lip}} \|\tilde{X}^N - X\|_p$; (ii) under Assumptions (A')-(B), $\|\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)\|_p = O(N^{-1/d})$ as $N \rightarrow \infty$.*

Lemma A.5. *Let Assumption (B) hold. For any $x \in S_X$, let $L(x) = 1/f^{Y|X=x}(q_\alpha(x))$ and $L^N(x) = 1/f^{Y|X=x}(c^N(x))$, where $c^N(x)$ is the infimum of all c 's between $\tilde{q}_\alpha^N(x)$ and $q_\alpha(x)$ for which*

$$\int_{\min(q_\alpha(x), \tilde{q}_\alpha^N(x))}^{\max(q_\alpha(x), \tilde{q}_\alpha^N(x))} f^{Y|X=x}(y) dy = f^{Y|X=x}(c) |\tilde{q}_\alpha^N(x) - q_\alpha(x)| \quad (\text{A.8})$$

(existence follows from the mean value theorem). Then $\|L^N(X)\|_p \rightarrow \|L(X)\|_p$ as $N \rightarrow \infty$.

Proof of Lemma A.3. Part (ii) of the result readily follows from Part (i) and Corollary 2.2, so that we may focus on the proof of Part (i). Note that $G_a(\tilde{X}^N)$ stands for the conditional

expectation of $\rho_\alpha(Y - a)$ given that $X = \tilde{X}^N$, which is different from $E[\rho_\alpha(Y - a)|\tilde{X}^N]$. For any a ,

$$\begin{aligned} |\tilde{G}_a(\tilde{X}^N) - G_a(X)| &\leq |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)| + |G_a(\tilde{X}^N) - G_a(X)| \\ &\leq \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)| + \sup_a |G_a(\tilde{X}^N) - G_a(X)| \end{aligned}$$

almost surely, so that

$$\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)| \leq \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)| + \sup_a |G_a(\tilde{X}^N) - G_a(X)|$$

almost surely. The triangular inequality then yields

$$\|\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)|\|_p \leq \|\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)|\|_p + \|\sup_a |G_a(\tilde{X}^N) - G_a(X)|\|_p. \quad (\text{A.9})$$

Since \tilde{X}^N is X -measurable, we have that

$$\tilde{G}_a(\tilde{X}^N) = E[\rho_\alpha(Y - a)|\tilde{X}^N] = E[E[\rho_\alpha(Y - a)|X]|\tilde{X}^N] = E[G_a(X)|\tilde{X}^N],$$

which yields

$$\begin{aligned} \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)| &= \sup_a |E[G_a(X) - G_a(\tilde{X}^N)|\tilde{X}^N]| \\ &\leq E\left[\sup_a |G_a(X) - G_a(\tilde{X}^N)| \middle| \tilde{X}^N\right], \end{aligned}$$

almost surely. From Jensen's inequality, we then obtain

$$\|\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(\tilde{X}^N)|\|_p \leq \|\sup_a |G_a(X) - G_a(\tilde{X}^N)|\|_p.$$

Substituting in (A.9) and using Lemma A.1(ii) yields

$$\begin{aligned} \|\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)|\|_p &\leq 2\|\sup_a |G_a(X) - G_a(\tilde{X}^N)|\|_p \\ &\leq 2\max(\alpha, 1 - \alpha)[m]_{\text{Lip}}\|\tilde{X}^N - X\|_p, \end{aligned}$$

which establishes the result. \square

Proof of Lemma A.4. Letting $\mathbb{I}_+ = \mathbb{I}_{[\tilde{G}_{\tilde{q}}(\tilde{X}^N) \geq G_q(X)]}$, note that

$$\begin{aligned} |\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)|\mathbb{I}_+ &\leq (\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X))\mathbb{I}_+ \\ &\leq (\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X))\mathbb{I}_+ \leq \left(\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)|\right)\mathbb{I}_+, \end{aligned}$$

almost surely. Similarly, with $\mathbb{I}_- = \mathbb{I}_{[\tilde{G}_{\tilde{q}}(\tilde{X}^N) < G_q(X)]}$, we have

$$\begin{aligned} |\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)|\mathbb{I}_- &\leq (G_q(X) - \tilde{G}_{\tilde{q}}(\tilde{X}^N))\mathbb{I}_- \\ &\leq (G_{\tilde{q}}(X) - \tilde{G}_{\tilde{q}}(\tilde{X}^N))\mathbb{I}_- \leq \left(\sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)|\right)\mathbb{I}_-, \end{aligned}$$

almost surely, so that $|\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)| \leq \sup_a |\tilde{G}_a(\tilde{X}^N) - G_a(X)|$, almost surely. The result then directly follows from Lemma A.3. \square

Proof of Lemma A.5. We have to prove that, as $N \rightarrow \infty$,

$$\int_{S_X} \frac{1}{(f^{Y|X=x}(c^N(x)))^p} dP_X(x) \rightarrow \int_{S_X} \frac{1}{(f^{Y|X=x}(q_\alpha(x)))^p} dP_X(x). \quad (\text{A.10})$$

First note that Assumption (B) ensures that, for any $x \in S_X$,

$$y \mapsto f^{Y|X=x}(y) = \frac{1}{m_2(x)} f^\varepsilon\left(\frac{y - m_1(x)}{m_2(x)}\right)$$

is continuous. Therefore, Theorem 3.2, which entails that $c^N(x) \rightarrow q_\alpha(x)$ for any x as $N \rightarrow \infty$, implies that

$$\frac{1}{(f^{Y|X=x}(c^N(x)))^p} \rightarrow \frac{1}{(f^{Y|X=x}(q_\alpha(x)))^p},$$

still for any x as $N \rightarrow \infty$. To establish (A.10), it is then sufficient — in view of Lebesgue's dominated convergence theorem — to prove that, for any x and any (sufficiently large) N ,

$$\frac{1}{f^{Y|X=x}(c^N(x))} = \frac{m_2(x)}{f^\varepsilon\left(\frac{c^N(x) - m_1(x)}{m_2(x)}\right)} \leq C \quad (\text{A.11})$$

for some constant C that does not depend on N .

To show (A.11), note that Theorem 3.2 and the continuity of $m_1(\cdot)$ and $m_2(\cdot)$ (with $m_2(\cdot)$ taking strictly positive values) over the compact set S_X entails that, for N sufficiently large,

$$\begin{aligned} \left| \frac{c^N(x) - m_1(x)}{m_2(x)} \right| &= \left| \frac{(q_\alpha(x) - m_1(x)) + (c^N(x) - q_\alpha(x))}{m_2(x)} \right| \\ &= \left| \varepsilon_\alpha + \frac{c^N(x) - q_\alpha(x)}{m_2(x)} \right| \\ &\leq |\varepsilon_\alpha| + \frac{|c^N(x) - q_\alpha(x)|}{D_1} \leq |\varepsilon_\alpha| + \frac{1}{D_1} = D_2, \end{aligned}$$

for some constants D_1, D_2 (that do not depend on N), where ε_α denotes the α -quantile of ε . Consequently, (A.11) directly follows from the continuity of $m_2(\cdot)$ and $f^\varepsilon(\cdot) : \mathbb{R} \rightarrow \mathbb{R}_0^+$. \square

Finally, we prove Theorem 3.1.

Proof of Theorem 3.1. Throughout the proof, we write $q(x)$ and $\tilde{q}(x)$ for $q_\alpha(x)$ and $\tilde{q}_\alpha^N(x)$, respectively. Similarly, q and \tilde{q} will stand for $q_\alpha(X)$ and $\tilde{q}_\alpha^N(X)$, respectively.

(i) Let first $r, s \in \mathbb{R}$ with $r \leq s$. For all $y \in \mathbb{R}$, one has $1 \geq \mathbb{I}_{[y \leq r]} + \mathbb{I}_{[y > s]}$, which implies

$$\rho_\alpha(y - r) - \rho_\alpha(y - s) \geq -(1 - \alpha)(s - r)\mathbb{I}_{[y \leq r]} + \alpha(s - r)\mathbb{I}_{[y > s]},$$

hence

$$\{\rho_\alpha(Y - \tilde{q}) - \rho_\alpha(Y - q)\}\mathbb{I}_{[\tilde{q} \leq q]} \geq \{-(1 - \alpha)(q - \tilde{q})\mathbb{I}_{[Y \leq \tilde{q}]} + \alpha(q - \tilde{q})\mathbb{I}_{[Y > q]}\}\mathbb{I}_{[\tilde{q} \leq q]}.$$

Taking expectation conditional on X , this gives

$$\begin{aligned}
|G_{\tilde{q}}(X) - G_q(X)|\mathbb{I}_{[\tilde{q} \leq q]} &= (G_{\tilde{q}}(X) - G_q(X))\mathbb{I}_{[\tilde{q} \leq q]} \\
&\geq \left\{ -(1-\alpha)(q - \tilde{q})P[Y \leq \tilde{q}|X] + \alpha(q - \tilde{q})P[Y > q|X] \right\}\mathbb{I}_{[\tilde{q} \leq q]} \\
&= (1-\alpha)(q - \tilde{q})(\alpha - P[Y \leq \tilde{q}|X])\mathbb{I}_{[\tilde{q} \leq q]} \\
&\geq \min(\alpha, 1-\alpha)|\tilde{q} - q|(P[Y \leq q|X] - P[Y \leq \tilde{q}|X])\mathbb{I}_{[\tilde{q} \leq q]} \\
&= \min(\alpha, 1-\alpha)|\tilde{q} - q|P[\min(\tilde{q}, q) < Y \leq \max(\tilde{q}, q)|X]\mathbb{I}_{[\tilde{q} \leq q]}, \tag{A.12}
\end{aligned}$$

almost surely.

Now, let $r, s \in \mathbb{R}$ with $r > s$. For all $y \in \mathbb{R}$, one has $1 \geq \mathbb{I}_{[y \leq s]} + \mathbb{I}_{[y > r]}$, which implies

$$\rho_\alpha(y - r) - \rho_\alpha(y - s) \geq -(1-\alpha)(s - r)\mathbb{I}_{[y \leq s]} + \alpha(s - r)\mathbb{I}_{[y > r]},$$

hence

$$\left\{ \rho_\alpha(Y - \tilde{q}) - \rho_\alpha(Y - q) \right\}\mathbb{I}_{[\tilde{q} > q]} \geq \left\{ -(1-\alpha)(q - \tilde{q})\mathbb{I}_{[Y \leq q]} + \alpha(q - \tilde{q})\mathbb{I}_{[Y > q]} \right\}\mathbb{I}_{[\tilde{q} > q]}.$$

Taking expectation conditional on X , this gives

$$\begin{aligned}
|G_{\tilde{q}}(X) - G_q(X)|\mathbb{I}_{[\tilde{q} > q]} &= (G_{\tilde{q}}(X) - G_q(X))\mathbb{I}_{[\tilde{q} > q]} \\
&\geq \left\{ -(1-\alpha)(q - \tilde{q})P[Y \leq q|X] + \alpha(q - \tilde{q})P[Y > q|X] \right\}\mathbb{I}_{[\tilde{q} > q]} \\
&= \alpha(q - \tilde{q})(P[Y > q|X] - (1-\alpha))\mathbb{I}_{[\tilde{q} > q]} \\
&\geq \min(\alpha, 1-\alpha)|\tilde{q} - q|(P[Y \leq \tilde{q}|X] - P[Y \leq q|X])\mathbb{I}_{[\tilde{q} > q]} \\
&= \min(\alpha, 1-\alpha)|\tilde{q} - q|P[\min(\tilde{q}, q) < Y \leq \max(\tilde{q}, q)|X]\mathbb{I}_{[\tilde{q} > q]}, \tag{A.13}
\end{aligned}$$

almost surely. Adding up (A.12) and (A.13) then provides

$$|G_{\tilde{q}}(X) - G_q(X)| \geq \min(\alpha, 1-\alpha)|\tilde{q} - q|P[\min(\tilde{q}, q) < Y \leq \max(\tilde{q}, q)|X]. \tag{A.14}$$

Now, for any $x \in S_X$,

$$\begin{aligned}
P[\min(\tilde{q}, q) < Y \leq \max(\tilde{q}, q)|X = x] &= \int_{\min(q(x), \tilde{q}(x))}^{\max(q(x), \tilde{q}(x))} f^{Y|X=x}(y) dy \\
&= f^{Y|X=x}(c^N(x)) |\tilde{q}(x) - q(x)| = \frac{|\tilde{q}(x) - q(x)|}{L^N(x)},
\end{aligned}$$

where $c^N(x)$ and $L^N(x)$ were defined in Lemma A.5, so that

$$P[\min(\tilde{q}, q) \leq Y < \max(\tilde{q}, q)|X] = \frac{|\tilde{q} - q|}{L^N(X)}$$

almost surely. Plugging into (A.14) yields

$$|\tilde{q} - q|^2 \leq \frac{1}{\min(\alpha, 1 - \alpha)} L^N(X) |G_{\tilde{q}}(X) - G_q(X)|,$$

or equivalently,

$$|\tilde{q} - q|^p \leq \frac{1}{(\min(\alpha, 1 - \alpha))^{p/2}} (L^N(X))^{p/2} |G_{\tilde{q}}(X) - G_q(X)|^{p/2}.$$

Taking expectations, applying Cauchy-Schwarz inequality in the righthand side, then computing p th roots, provides

$$\|\tilde{q} - q\|_p \leq \frac{1}{\sqrt{\min(\alpha, 1 - \alpha)}} \|L^N(X)\|_p^{1/2} \|G_{\tilde{q}}(X) - G_q(X)\|_p^{1/2}. \quad (\text{A.15})$$

From Lemmas A.3-A.4, we obtain

$$\begin{aligned} \|G_{\tilde{q}}(X) - G_q(X)\|_p &\leq \|G_{\tilde{q}}(X) - \tilde{G}_{\tilde{q}}(\tilde{X}^N)\|_p + \|\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)\|_p \\ &\leq \|\sup_a |G_a(X) - \tilde{G}_a(\tilde{X}^N)|\|_p + \|\tilde{G}_{\tilde{q}}(\tilde{X}^N) - G_q(X)\|_p \\ &\leq 4 \max(\alpha, 1 - \alpha) [m]_{\text{Lip}} \|\tilde{X}^N - X\|_p. \end{aligned}$$

The result then follows by plugging this into (A.15) (the boundedness of $L^N(X)$ in L_p is a direct corollary of Lemma A.5).

(ii) The result directly follows from Part (i) and Corollary 2.2. \square

B Proof of Theorem 4.1

Let $\gamma^N = \gamma^N(X) = \{\tilde{x}_1, \dots, \tilde{x}_N\}$ be an optimal grid and $\hat{\gamma}^{N,n} = \hat{\gamma}^{N,n}(X_1, \dots, X_n) = (\hat{x}_1^{N,n}, \dots, \hat{x}_N^{N,n})$ be the grid provided by the CLVQ algorithm. Throughout this section, we assume the almost sure convergence of the empirical quantization of X to the population one, that is

$$\hat{X}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(X) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \text{Proj}_{\gamma^N}(X) = \tilde{X}^N, \quad (\text{B.1})$$

which is justified by the discussion in Section 2.3.

The proof of Theorem 4.1 then requires Lemmas B.1-B.2 below.

Lemma B.1. *Let Assumption (C) hold. Fix $N \in \mathbb{N}_0$ and $x \in S_X$, and write $\tilde{x} = \text{Proj}_{\gamma^N}(x)$ and $\tilde{x} = \text{Proj}_{\hat{\gamma}^{N,n}}(x)$. Then, with $\hat{X}_i^N = \text{Proj}_{\hat{\gamma}^{N,n}}(X_i)$, $i = 1, \dots, n$, we have*

- (i) $\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[\hat{X}_i^N = \tilde{x}^N]} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} P[\tilde{X}^N = \tilde{x}];$
- (ii) *after possibly reordering the \tilde{x}_i 's, $\hat{x}_i^{N,n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \tilde{x}_i$, $i = 1, \dots, N$ (hence, $\hat{\gamma}^{N,n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \gamma^N$).*

Proof. Under (B.1), Part (i) was shown in Bally et al. (2005) (see also Pagès (1998)) and Part (ii) only states the a.s. convergence of the supports $\hat{\gamma}^{N,n}$ to γ^N , which is a necessary condition for the corresponding convergence of random vectors in (B.1). \square

Lemma B.2. Fix $\alpha \in (0, 1)$, $x \in S_X$ and $N \in \mathbb{N}_0$, let $K \subset \mathbb{R}$ be compact, and define

$$\widehat{G}_a^{N,n}(\hat{x}^N) := \frac{\frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}}{\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]}}.$$

Then, under Assumptions (A) and (C), (i) $\sup_{a \in K} |\widehat{G}_a^{N,n}(\hat{x}^N) - \widetilde{G}_a(\tilde{x})| = o_P(1)$ as $n \rightarrow \infty$; (ii) $|\min_{a \in \mathbb{R}} \widehat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \widetilde{G}_a(\tilde{x})| = o_P(1)$ as $n \rightarrow \infty$; (iii) $|\widetilde{G}_{\widehat{q}_\alpha^{N,n}(x)} - \widetilde{G}_{\widetilde{q}_\alpha^N(x)}| = o_P(1)$ as $n \rightarrow \infty$.

Proof. (i) Since

$$\widetilde{G}_a(\tilde{x}) = \mathbb{E}[\rho_\alpha(Y - a) | \widetilde{X}^N = \tilde{x}] = \frac{\mathbb{E}[\rho_\alpha(Y - a) \mathbb{I}_{[\widetilde{X}^N = \tilde{x}]}]}{P[\widetilde{X}^N = \tilde{x}]},$$

it is sufficient — in view of Lemma B.1(i) — to prove that, as $n \rightarrow \infty$,

$$\sup_{a \in K} \left| \frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]} - \mathbb{E}[\rho_\alpha(Y - a) \mathbb{I}_{[\widetilde{X}^N = \tilde{x}]}] \right| = o_P(1).$$

Of course, it is natural to consider the decomposition

$$\sup_{a \in K} \left| \frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]} - \mathbb{E}[\rho_\alpha(Y - a) \mathbb{I}_{[\widetilde{X}^N = \tilde{x}]}] \right| \leq \sup_{a \in K} |T_{a1}| + \sup_{a \in K} |T_{a2}|,$$

with

$$T_{a1} = \frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - a) (\mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]} - \mathbb{I}_{[\widetilde{X}_i^N = \tilde{x}]})$$

and

$$T_{a2} = \frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{I}_{[\widetilde{X}_i^N = \tilde{x}]} - \mathbb{E}[\rho_\alpha(Y - a) \mathbb{I}_{[\widetilde{X}^N = \tilde{x}]}].$$

Using the fact that $m_1(\cdot)$ and $m_2(\cdot)$ are continuous functions defined over the compact set S_X , we obtain that, for any $a \in K$, there exist positive constants C_1 and C_2 such that

$$\rho_\alpha(Y - a) \leq \max(\alpha, 1 - \alpha) |Y - a| \leq \max(\alpha, 1 - \alpha) (|m_1(X)| + |m_2(X)| |\varepsilon| + |a|) \leq C_1 + C_2 |\varepsilon|,$$

that is in L_1 (recall that ε is assumed to be in L_p , $p = 2$), the uniform law of large numbers (see, e.g., Theorem 16(a) in Ferguson, 1996) shows that $\sup_{a \in K} |T_{a2}| = o_P(1)$ as $n \rightarrow \infty$.

Turning to T_{a1} , consider the set $\mathcal{I}_n = \{i = 1, \dots, n : \mathbb{I}_{[\widehat{X}_i^N = \hat{x}^N]} \neq \mathbb{I}_{[\widetilde{X}_i^N = \tilde{x}]}\}$ collecting the indices of observations that are projected on the same point as x for γ^N but not for $\hat{\gamma}^{N,n}$ (or

vice versa on the same point as x for $\hat{\gamma}^{N,n}$ but not for γ^N). For any $a \in K$, we have

$$\begin{aligned} |T_{a1}| &\leq \frac{1}{n} \sum_{i \in \mathcal{I}_n} |\rho_\alpha(Y_i - a)| \leq \frac{\max(\alpha, 1 - \alpha)}{n} \sum_{i \in \mathcal{I}_n} (|m_1(X_i)| + |m_2(X_i)| |\varepsilon_i| + |a|) \\ &\leq \frac{\#\mathcal{I}_n}{n} \times \frac{1}{\#\mathcal{I}_n} \sum_{i \in \mathcal{I}_n} (C_1 + C_2 |\varepsilon_i|) =: S_1 \times S_2. \end{aligned}$$

Clearly, Lemma B.1(ii) implies that $\#\mathcal{I}_n/n = o_P(1)$ as $n \rightarrow \infty$, while the independence between \mathcal{I}_n (which is measurable with respect to the X_i 's) and the ε_i 's entails that $E[S_2] = O(1)$ as $n \rightarrow \infty$, so that S_2 is bounded in probability. Consequently, $\sup_{a \in K} |T_{a1}|$ goes to zero in probability as $n \rightarrow \infty$. Part (i) of the result follows.

(ii) Fix $\delta > 0$ and $\eta > 0$. Writing $\hat{q} = \hat{q}_\alpha^{N,n}(x)$ and, as in the previous section, $\tilde{q} = \tilde{q}_\alpha^N(x)$, first choose n_1 and M large enough to have $|\tilde{q}| \leq M$ and $P[|\hat{q}| > M] < \eta/2$ for any $n \geq n_1$ (Lemma B.1(i) implies that \hat{q} is the sample quantile of a number of Y_i 's that increases to infinity, so that $|\hat{q}|$, with arbitrarily large probability for n large, cannot exceed $2 \sup_{x \in S_X} |q_\alpha(x)|$). Then, with $\mathbb{I}_+ = \mathbb{I}_{[\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) \geq \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})]}$, we have

$$\begin{aligned} |\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})| \mathbb{I}_+ &= (\hat{G}_{\hat{q}}^{N,n}(\hat{x}^N) - \tilde{G}_{\tilde{q}}(\tilde{x})) \mathbb{I}_+ \\ &\leq (\hat{G}_{\tilde{q}}^{N,n}(\hat{x}^N) - \tilde{G}_{\tilde{q}}(\tilde{x})) \mathbb{I}_+ \leq \sup_{a \in [-M, M]} |\hat{G}_a^{N,n}(\hat{x}^N) - \tilde{G}_a(\tilde{x})| \mathbb{I}_+, \end{aligned} \quad (\text{B.2})$$

almost surely. Now, with $\mathbb{I}_- = \mathbb{I}_{[\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) < \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})]}$, we have that, under $|\hat{q}| \leq M$,

$$\begin{aligned} |\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})| \mathbb{I}_- &= |(\tilde{G}_{\hat{q}}(\tilde{x}) - \hat{G}_{\hat{q}}^{N,n}(\hat{x}^N))| \mathbb{I}_- \\ &\leq |(\tilde{G}_{\hat{q}}(\tilde{x}) - \hat{G}_{\hat{q}}^{N,n}(\hat{x}^N))| \mathbb{I}_- \leq \sup_{a \in [-M, M]} |\hat{G}_a^{N,n}(\hat{x}^N) - \tilde{G}_a(\tilde{x})| \mathbb{I}_-. \end{aligned} \quad (\text{B.3})$$

By combining (B.2) and (B.3), we obtain that, under $|\hat{q}| \leq M$,

$$|\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})| \leq \sup_{a \in [-M, M]} |\hat{G}_a^{N,n}(\hat{x}^N) - \tilde{G}_a(\tilde{x})|.$$

Consequently, for any $n \geq n_1$, we obtain

$$\begin{aligned} P \left[\left| \min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) \right| > \delta \right] &= P \left[\left| \min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) \right| > \delta, |\hat{q}| \leq M \right] + P \left[\left| \min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) \right| > \delta, |\hat{q}| > M \right] \\ &\leq P \left[\sup_{a \in [-M, M]} |\hat{G}_a^{N,n}(\hat{x}^N) - \tilde{G}_a(\tilde{x})| > \delta \right] + \frac{\eta}{2}. \end{aligned}$$

From Part (i) of the lemma, the first term is smaller than $\eta/2$ for any $n \geq n_2$. We conclude that, for any $n \geq n_0 := \max(n_1, n_2)$, we have

$$P \left[\left| \min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x}) \right| > \delta \right] < \eta,$$

which shows Part (ii) of the result.

(iii) The proof proceeds in the same way as in (ii) above. First we pick n_1 and M large enough to have $P[|\hat{q}| > M] < \eta/2$ for any $n \geq n_1$, which yields

$$P\left[|\tilde{G}_{\hat{q}}(\tilde{x}) - \tilde{G}_{\hat{q}}(\tilde{x})| > \delta\right] \leq P\left[|\tilde{G}_{\hat{q}}(\tilde{x}) - \tilde{G}_{\hat{q}}(\tilde{x})| > \delta, |\hat{q}| \leq M\right] + \frac{\eta}{2}. \quad (\text{B.4})$$

Now, from the triangular inequality, we obtain

$$\begin{aligned} & P\left[|\tilde{G}_{\hat{q}}(\tilde{x}) - \tilde{G}_{\hat{q}}(\tilde{x})| > \delta, |\hat{q}| \leq M\right] \\ & \leq P\left[|\tilde{G}_{\hat{q}}(\tilde{x}) - \hat{G}_{\hat{q}}^{N,n}(\hat{x}^N)| > \delta/2, |\hat{q}| \leq M\right] + P\left[|\hat{G}_{\hat{q}}^{N,n}(\hat{x}^N) - \tilde{G}_{\hat{q}}(\tilde{x})| > \delta/2, |\hat{q}| \leq M\right] \\ & \leq P\left[|\min_{a \in \mathbb{R}} \hat{G}_a^{N,n}(\hat{x}^N) - \min_{a \in \mathbb{R}} \tilde{G}_a(\tilde{x})| > \delta/2\right] + P\left[\sup_{a \in [-M, M]} |\hat{G}_a^{N,n}(\hat{x}^N) - \tilde{G}_a(\tilde{x})| > \delta/2\right], \end{aligned}$$

which, from Part (i) and Part (ii) of the lemma, can be made arbitrarily small for n large enough. Jointly with (B.4), this establishes the result. \square

We can now conclude with the proof of Theorem 4.1.

Proof of Theorem 4.1. Since the function $\rho_\alpha(\cdot)$ is strictly convex, $\tilde{G}_a(\tilde{x})$ is also strictly convex in a . Its minimum in a (for any fixed x) is therefore unique, and the convergence in probability of $\tilde{G}_{\hat{q}}(\tilde{x})$ towards $\tilde{G}_{\hat{q}}(\tilde{x})$ implies the convergence in probability of the corresponding arguments. \square

References

- Azaïs, R., A. Gégout-Petit, and J. Saracco (2012). Optimal quantization applied to sliced inverse regression. *J. Statist. Plann. Inference* 142(2), 481–492.
- Bally, V., G. Pagès, and J. Printems (2005). A quantization tree method for pricing and hedging multidimensional American options. *Math. Finance* 15(1), 119–168.
- Bhattacharya, P. K. and A. K. Gangopadhyay (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist.* 18(3), 1400–1415.
- Charlier, I., D. Paindaveine, and J. Saracco (2014). Numerical study of a conditional quantile estimator based on optimal quantization. *Manuscript in preparation*.
- Fan, J., T.-C. Hu, and Y. Truong (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics* 21(4), 433–446.
- Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall/CRC.
- Fischer, A. (2010). Quantization and clustering with bregman divergences. *Journal of Multivariate Analysis* 101, 2207–2221.

- Fischer, A. (2013). Présentation de deux méthodes d'apprentissage non supervisé : quantification et courbes principales. *Soumis*.
- Gannoun, A., S. Girard, C. Guinot, and J. Saracco (2002). Reference curves based on non-parametric quantile regression. *Statistics in Medicine* 21(4), 3119–3135.
- Graf, S. and H. Luschgy (2000). *Foundations of quantization for probability distributions*, Volume 1730 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag.
- Heagerty, P. and M. Pepe (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. *J. R. Stat. Soc. Ser. C Appl. Stat.* 48(4), 533–551.
- Koenker, R. and G. Bassett, Jr. (1978). Regression quantiles. *Econometrica* 46(1), 33–50.
- Pagès, G. (1998). A space quantization method for numerical integration. *J. Comput. Appl. Math.* 89(1), 1–38.
- Pagès, G., H. Pham, and J. Printems (2004a). An optimal Markovian quantization algorithm for multi-dimensional stochastic control problems. *Stoch. Dyn.* 4(4), 501–545.
- Pagès, G., H. Pham, and J. Printems (2004b). Optimal quantization methods and applications to numerical problems in finance. In *Handbook of computational and numerical methods in finance*, pp. 253–297. Boston, MA: Birkhäuser Boston.
- Pagès, G. and J. Printems (2003). Optimal quadratic quantization for numerics: the Gaussian case. *Monte Carlo Methods Appl.* 9(2), 135–165.
- Yu, K. and M. C. Jones (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* 93(441), 228–237.
- Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *J. R. Stat. Soc. Ser. D Statistician* 52(3), 331–350.
- Zador, P. L. (1964). *Development and evaluation of procedures for quantizing multivariate distributions*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.